Journal of the Formosan Medical Association (2014) 113, 179–186

Available online at www.sciencedirect.com

SciVerse ScienceDirect

journal homepage: www.jfma-online.com

ORIGINAL ARTICLE

Likelihood ratios of multiple cutoff points of the Taipei City Developmental Checklist for Preschoolers, 2nd version

Hua-Fang Liao [a,b,*], Grace Yao [c], Cheng-Chun Chien [c], Ling-Yee Cheng [d], Wu-Shiun Hsieh [e]

[a] *School and Graduate Institute of Physical Therapy, College of Medicine, National Taiwan University, Taipei, Taiwan*
[b] *Department of Rehabilitation Medicine, National Taiwan University Hospital, Taipei, Taiwan*
[c] *Department of Psychology, National Taiwan University, Taipei, Taiwan*
[d] *Department of Physical Medicine and Rehabilitation, Taipei Veterans General Hospital, Taipei, Taiwan*
[e] *Department of Pediatrics, National Taiwan University Hospital and National Taiwan University, College of Medicine, Taipei, Taiwan*

Received 15 September 2010; received in revised form 11 July 2011; accepted 4 October 2011

KEYWORDS
child;
decision making;
developmental disabilities;
reliability and validity

Background/Purpose: This study aimed to examine the reliability and clinical decision validities of the Taipei City Developmental Checklist for Preschoolers, 2nd version (the Taipei II, which was filled out by parents) and the screening procedures conducted in the medical setting.
Methods: Methodology research and case control study designs were adopted. A total of 310 dyads consisting of children who were developing typically and 196 dyads of children with developmental delays and age 5.5 to 35.5 months were recruited for validity test. Among them, 165 mothers filled out the questionnaire twice within 1 week to examine the test–retest reliability of the total score and individual items. Validity indexes of the single cutoff strategy and multiple cutoff strategies were analyzed. With two cutoff point strategies, the likelihood ratios (LR) of the three test results, positive, neutral, and negative, were calculated.
Results: The test–retest reliabilities of the total scores of the seven checklists of the Taipei II (r_s = 0.54–0.89, $p<0.05$) and their individual items (agreement 92% to 100%) were acceptable, except for the 30-month checklist and three individual items. The positive LR (LR+) and negative LR (LR-) of the single cutoff strategy were acceptable with most LR+ more than 2, and all LR- less than 0.5. Most of the diagnostic odds ratios of single cutoff strategies were less than 50 and they did not meet the acceptable criteria. When multiple cutoff points were used, all of the LRs with

Conflicts of interest: The authors have no conflicts of interest relevant to this article.

* Corresponding author. School and Graduate Institute of Physical Therapy, National Taiwan University College of Medicine, Third floor, 17 Xuzhou Road, Taipei 100, Taiwan.
E-mail address: hfliao@ntu.edu.tw (H.-F. Liao).

doi:10.1016/j.jfma.2011.10.005

positive test results were equal to infinity that met SpPin criteria, and all of the LRs with negative test results less than 0.5 had at least a small but important diagnostic impact.
Conclusion: Taipei II with multiple cutoff points could give more useful clinical information than using a single cutoff point. The multiple likelihood ratios of Taipei II for children older than 3 years and in different cultural backgrounds need further study.

Introduction

The benefits of early intervention for toddlers with developmental delays (DD) have been shown in randomized controlled trials.[1] Therefore, it is important that reliable and valid screening tests be administered earlier to avoid unreliable recall of milestones and the underdetection of clinical judgment (as in clinics, diagnoses of developmental delays based on clinical vignettes only could be misguided).[2,3]

A developmental screening test covering various developmental domains and with proper cutoff points of sound validity is helpful to detect children with DD earlier and correctly.[4,5] The Taipei City Developmental Checklist for Preschoolers, 2nd version (Taipei II), revised in 2005, is a concise screening instrument that aims to identify children who should receive further assessment due to the potential risks of developmental delays or disabilities. It has been applied widely in Taiwan in recent years[6,7] and it has four language versions: traditional Chinese, Indonesian, Thai, and Vietnamese.

The psychometric properties of previous studies of this test as completed by parent-targeted questionnaires are unknown in a medical setting, as testing has been carried out mostly in community settings and conducted by clinical psychologists. According to the screening policy in Taiwan proposed by the Department of Health, developmental surveillance of each infant and toddler is conducted six times before age 3 years to fit the vaccination schedule in medical settings. Involving parents in the assessment and intervention can enhance their knowledge of child development[8] and is cost-efficient.[9] Therefore, it is necessary to reexamine the validities of decisions made based on the Taipei II used in the medical setting and filled out by parents.

Validation is a key step in the development of the suggested cutoff point. Validation ideally is investigated on a group of children distinct from the group used to develop it.[10] The developmental surveillance of each child is conducted six times before age 3 in Taiwan. Therefore, the authors recruited a group of toddlers to examine the validity of the suggested cutoff strategies of the Taipei II. As multilevel likelihood ratios of a test with multiple cutoff points are more powerful and useful than one single cutoff point,[11] the purposes of this study were to investigate the test–retest reliability, its validity for decision making, and the multilevel likelihood ratios of the Taipei II in a medical setting for infants and toddlers less than age 36 months.

Materials and methods

Participants

We recruited dyads from two medical centers, one local hospital, and one developmental assessment center in Taipei City, as well as one local pediatric clinic in Chiayi City. Children with developmental delays were diagnosed as having developmental delays or developmental disabilities by a developmental assessment team and then referred for early intervention. Children developing typically were free from any neuromuscular, musculoskeletal, or cardiopulmonary disease. The children were ascertained by pediatricians as being developmentally typical after taking their histories, conducting physical examinations, using the developmental surveillance items of the Child Health Pamphlet,[12] and conducting a chart review at the well baby clinics. Parents signed a consent form that was reviewed and approved by the Institutional Review Board of one medical center.

A total of 310 dyads comprised of children developing typically (DT), and 196 dyads comprised of children having DD, were recruited for validity testing. Their ages ranged from 5.5 months to 35.5 months and the demographic data are shown in Table 1. Among the DT children, 165 mothers filled out the questionnaire twice within 1 week to examine the test–retest reliability of the Taipei II. Table 4 shows the numbers of the two groups divided into the Taipei II's seven age group checklists (6, 9, 12, 15, 18, 24 and 30 months).

Measurement

The Taipei II provides 13 checklists for 13 age groups: 4, 6, 9, 12, 15, 18, 24, 30, 36, 42, 48, 60, and 72 months. Each

Table 1 Basic data of children developing typically or with developmental delays.

	Developing typically (n = 310)	Developmental delay (n = 196)
Child characteristics		
Age of children, mean ± SD (months)*	16.7 ± 8.3	21.6 ± 8.1
Male sex, n (%)*	140 (45%)	124 (63%)
Premature, n (%)*	68 (22%)	63 (33%)
Family characteristics		
Maternal age, mean ± SD (years)	32.7 ± 4.2	33.4 ± 4.5
Career mother, n (%)*	154 (50%)	55 (29%)
Taiwanese mother, n (%)	294 (95%)	179 (92%)
Maternal education < high school, n (%)	7 (2%)	9 (5%)
Paternal age, mean ± SD (years)	35.3 ± 4.9	36.0 ± 5.3

Note: there were missing data for some variables.
*Significant differences between DT and DD groups ($p<0.05$, by independent t-test or Chi-square test).

checklist lists 11 to 13 behaviors or skills related to gross motor, fine motor, cognition, language/communication, and emotion/social areas easily observed or elicited by the child's caregiver. The internal consistency coefficients (α) of the Taipei II's 13 checklists were 0.72–0.87.[6] A validity study of the Taipei II from a sample of 3,792 children age 4 months to 72 months in the community setting (n = 3,146) or medical care institutes (n = 646) showed that the sensitivity ranged from 0.85–1.00 and specificity was 0.82–1.00 for cutoff strategy A. Sensitivity for cutoff strategy B ranged from 0.75–1.00, and specificity was 0.72–1.00.[7] Cut-off strategy A was set at ≥1 item failure, while cut-off strategy B was set at ≥2 items or ≥1 marked item failure.

Procedure

For the reliability and validity of the Taipei II, methodology research and case control study designs were adopted. After parents signed the consent form, demographic data of their children were collected. To simulate the clinical situation, after explaining the purpose of this study and the rating principles, the Taipei II checklist was filled out by one of the parents or main caregivers at clinics for validity analysis. To examine the test–retest reliability, the data of the Taipei II were collected twice within a time interval of 1 week. The trained tester would answer any queries proposed by the parents without further hints.

For children age 5.51–8.50 months, 8.51–11.50 months, 11.51–14.50 months, 14.51–17.50 months, 17.51–23.50 months, 23.51–29.50 months, or 29.51–35.50 months, their parents or main caregivers filled out the 6-, 9-, 12-, 15-, 18-, 24- or 30-month checklist of the Taipei II, respectively. Diagnostic data of development delays or disabilities were collected from the medical records or filled out by pediatricians.

Data analysis

In the Taipei II, each checklist has some positive statement items and some negative statement items. The positive statement items are the behaviors or skills expected to be achieved at that age. The negative statement items are the expected behaviors or skills not achieved or observed, or deviated behaviors usually not observed in DT children. All items are category scales. The respondent has to answer "yes" or "no" for each item. Answering yes for each positive statement item, or no for each negative statement item, would be scored 1. For data analysis, a total score of each age appropriate checklist was calculated. Therefore, the range of total scores was from 0 to 11–13, depending on the number of items on each checklist. The higher the total score, the less the probability of developmental delay.

For test–retest reliability, we analyzed the total score as well as the individual item, because item-level reliability can provide information about whether an item needs to be revised or replaced. The values of the Shapiro-Wilk test and the Kolmogorov-Smirnov test were used to examine the distribution of data in this study. Most data were against the normal distribution assumption. Therefore, the nonparametric test, Spearman correlation was used to examine the test–retest reliability.

For clinical application, we calculated the decision validity indexes: sensitivity, specificity, the Youden index (YI), positive likelihood ratio (LR+), negative likelihood ratio (LR-), and the diagnostic odds ratio (DOR) for strategies A and B. A likelihood ratio is the likelihood of a given test result in a patient with the target disorder compared with the likelihood of the same result in a patient without that disorder.[13] The formula of the likelihood ratio for a positive test result is: LR+ = sensitivity/(1 − specificity). The formula of the likelihood ratio for a negative test result is: LR− = (1 − sensitivity)/specificity. In general, sensitivity levels of 70% or more are acceptable[14] in order to limit the number of false negatives.[15] Specificity levels of 70% to 80% are acceptable.[14] The probability of a disease after a test (posttest probability)—developmental delay in this study—usually is obtained by calculating the LR of the test used and using formulas based on Bayesian theorem or a nomogram.[11] However, some argued that such calculations are not necessary for tests with high sensitivity or high specificity.[10,11]

Negative results from highly sensitive tests can rule out a diagnosis (sensitive, negative, out = SnNout), and positive results from highly specific tests can rule in a diagnosis (specificity, positive, in = SpPin).[11] After examining some diagnostic test studies, Pewsner proposed that the power of a test to rule out or rule in a diagnosis depended on both sensitivity and specificity.[16] The likelihood ratios depend on both sensitivity and specificity. For the LR of infinite, the posttest probability of a positive test result must be very, very high, and it was defined as SpPin in this study. For the LR of 0, the posttest probability of a negative test result must be 0, and it was defined as SnNout in this study.

Likelihood ratios greater than 10 or less than 0.1 generate large and often conclusive changes from pretest to posttest probability. Likelihood ratios of 5 to 10 and 0.1 to 0.2 generate moderate shifts in pretest to posttest probability. Likelihood ratios of 2 to 5 and 0.5 to 0.2 generate small (but sometimes important) changes in probability. Likelihood ratios of 0.5 to 2 alter probability to a small (and rarely important) degree and they have an indeterminate diagnostic impact.[17]

The overall diagnostic indexes used in this study were the Youden index (YI), the diagnostic odds ratio (DOR), and the area under the receiver operating characteristics curve (AUC). The YI is calculated in the formula: YI = sensitivity (%) + specificity (%) − 100%. It is independent of prevalence. The larger the YI, the better the validity. A test with YI equal to 0 is a useless test.[10] The DOR of a test is the ratio of the odds of a positive result with disease, relative to the odds of the positive result without disease.[13] The DOR can be calculated by the formula as follows: DOR = (TP/FN)/(FP/TN) = (positive likelihood ratio)/(negative likelihood ratio).

The value of the DOR ranges from 0 to infinity. A higher DOR value means good separation between positive and negative test results. A DOR value less than 1 means improper test interpretation.[13] The minimum acceptable value of DOR is 50, and a value >500 is very good.[13] The 95% Confidence Interval (CI) of DOR was also calculated.[13]

The receiver operating characteristics curve (ROC) analysis is defined as a plot graph with test sensitivity as the y axis, and 1− specificity as the x axis. This is an effective method of evaluating the quality or performance of screening tests.[18,19] The AUC represents a single value that summarizes the discriminative ability of a test across the full range of cutoffs, and which is independent of prevalence. Perfect tests produce an AUC of 1.0. The area under the ROC curve greater than 0.9 has high accuracy; 0.7—0.9 indicates moderate accuracy, 0.5—0.7 is low accuracy, and 0.5 means a chance result.[18] All statistical analyses were performed by using the Statistical Package for Social Science version 13.0 (SPSS Inc., Chicago, Illinois, USA). The level of statistical significance in this study was set at $\alpha < 0.05$ for two-tailed tests.

Results

The basic data of the participants are shown in Table 1. There were no significant differences in the family characteristics between the two groups, except that the DT group had a higher proportion of career mothers than the DD group, and the DD group had a higher mean age, more males, and more premature babies than the DT group. The test—retest reliabilities of the total score and of each item of the Taipei II in seven age groups of DT children are shown in Table 2. Except for the 30-month checklist, the reliability coefficients of the total score of other checklists were above 0.5 and significant. Except for the third, forth and fifth items of the 6-month checklist, the third item of the 12-month checklist, and the ninth item of the 24-month checklist, all of the items had agreement more than 90%.

While using clinical diagnosis as the criteria, the validity indexes of the Taipei II in cutoff strategy A or B in different age groups are presented in Table 3. The values of sensitivity are 84%—100% and 67%—100% in strategies A and B, respectively, and of specificity 20%—70% and 49%—93% respectively. All sensitivities were above 70% except for strategy B of 9- and 18-month checklists. Less than half of the checklists had specificities above 70%. Those with specificities above 70% were strategy A of the 24-month checklist, and strategy B of the 9-, 15-, 18-, 24- and 30-month checklists. All YI were above 0 and ranged from 12% to 82%. Most of the values of LR+ were more than the minimal criteria of 2, except for strategy A of the 6- and 9-month checklists and both strategies of the 12-month checklist. The LR+ in strategy B of the 24-month checklist was the highest, 11.16. All of the values of LR- were less than the minimal criteria of 0.5, with that of the 15-month checklist being the lowest. For DOR, all checklists were above 1, and only that of the 15-month was higher than 50.

The results of the ROC curve analyses showed that there was significant and moderate to high screening accuracy ($p<0.05$) for each age appropriate checklist (6-, 9-, 12-, 15-, 18-, 24- and 30-month) with AUC (95% CI) of 0.85 (0.72—0.98); 0.72 (0.55—0.90); 0.81 (0.69—0.92); 0.96 (0.92—1.0); 0.84 (0.76—0.92); 0.90 (0.82—0.97); and 0.86 (0.77—0.94), respectively.

The multilevel likelihood ratios and the diagnostic impacts of two cutoff points on the Taipei II's seven checklists at different age groups are presented in Table 4. The chosen cutoff points were different in different age-appropriate checklists from the empirical data. In all checklists, the positive test results had very high likelihood ratios that met the SpPin criteria. The negative test results of the 15-month checklist had very low likelihood ratios that met the SnNout criteria. The values of negative likelihood ratios of other checklists that had a certain degree of diagnostic impact were less than 0.5.

Discussion

The results of this study showed that the test—retest reliabilities of the total scores of the seven checklists and their individual items of the Taipei II were acceptable, except for the total score of the 30-month checklist for children under age 3. All YI of both strategies were acceptable, although not high; all were above 0, which meet the minimal requirement. The LR+ and LR- of cutoff strategy A or B, the single cutoff strategy, were acceptable, with most values of LR+ more than 2 and all LR- less than 0.5. However, only one checklist had a DOR higher than 50. Most checklists did not meet the minimal acceptable DOR value. That might mean that most checklists were not good enough to separate positive and negative test results if only one cutoff point (either strategy A or B) was used.

Instead, when multiple cutoff points were used, the children who had positive results could be ruled in and may

Table 2 Test—retest reliabilities of total scores and each item of the Taipei II's seven age groups of children developing typically.

Age group	*n*	r_s of total score	Agreement (%) of individual item												
			1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	12th	13th
6 months	37	0.55‡	97	97	76	89	84	97	100	92	92	100	100	100	
9 months	21	0.89‡	100	100	95	100	100	95	100	100	100	100	100	96	
12 months	20	0.64†	100	100	85	90	95	100	95	100	100	80	100		
15 months	25	0.92‡	96	96	100	100	100	96	100	100	100	96	100	100	
18 months	27	0.76‡	100	93	100	100	100	100	96	100	100	100	93	100	100
24 months	16	0.54*	100	100	100	100	100	100	100	100	88	94	100		
30 months	19	0.19	90	100	100	100	100	100	100	100	100	95	95	95	

*$p<0.05$, †$p<0.01$, ‡$p<0.001$, by Spearman correlation test.

Table 3 Validity indexes of the Taipei II in cutoff strategy A or B using clinical diagnosis as the criteria.

Age group	Cutoff strategies*	Sensitivity (%)	Specificity (%)	Youden index (%)	Positive likelihood ratio	Negative likelihood ratio	DOR (95% CI)
6 months (n = 78)	A	93	47	40	1.75	0.15	11 (1—93)
	B	79	69	48	2.51	0.31	8 (2—32)
9 months (n = 47)	A	92	20	12	1.15	0.42	3 (0—25)
	B	67	71	38	2.33	0.47	5 (1—20)
12 months (n = 64)	A	95	35	30	1.46	0.14	11 (1—88)
	B	90	49	39	1.77	0.20	9 (2—44)
15 months (n = 64)	A	100	69	69	3.27	0	∞
	B	100	82	82	5.44	0	∞
18 months (n = 98)	A	85	64	49	2.37	0.23	10 (4—28)
	B	67	78	45	3.03	0.43	7 (3—17)
24 months (n = 73)	A	93	70	63	3.10	0.10	31 (8—127)
	B	74	93	68	11.16	0.27	41 (8—200)
30 months (n = 82)	A	84	69	53	2.72	0.24	12 (4—33)
	B	70	82	52	3.89	0.37	11 (4—30)
Whole group (n = 506)	A	90	54	43	1.93	0.19	10 (6—17)
	B	75	74	49	2.87	0.34	8 (6—13)

*Strategy A = number of failure items≥1; Strategy B = number of failure items ≥2 or failure star items≥1. DOR = diagnostic odds ratio.

be diagnosed as suspected developmental delays and referred for further comprehensive tests, examination, or intervention. Moreover, those children who had negative test results could be ruled out and diagnosed as within the normal range. Those children who had neutral test results could be referred for a second screening test or closely monitored in the next visit. Multilevel likelihood ratios with multiple cutoff points could have more useful clinical applications than the single cutoff point.[11] To the authors' knowledge, there is only one previous study using multiple cutoff points for developmental screening.[12]

Previous screening tests usually provide the decision validity index with a single cutoff point, such as sensitivity, specificity, LR+, LR-, etc.[4,8,9] However, there are trade-offs between sensitivity and specificity in different cutoff points.[20] Variations in diagnostic criteria, test setting, and target population also affect the sensitivity and specificity of a developmental test.[21] This study used a testing

Table 4 The likelihood ratio and the diagnostic impact of multiple cutoff points of seven age groups of the Taipei II.

Age group	Test results	Total score	DD (n)	DT (n)	Likelihood ratio	Diagnostic impact	AUC
6 months (n = 78)	Positive	<8	7	0	∞	SpPin, rule-in	0.85
	Neutral	8—11	6	34	0.81	Indeterminate	
	Negative	12	1	30	0.15	Moderate	
9 months (n = 47)	Positive	<9	1	0	∞	SpPin, rule-in	0.72
	Neutral	9—11	10	28	1.04	Indeterminate	
	Negative	12	1	7	0.42	Small	
12 months (n = 64)	Positive	<5	1	0	∞	SpPin, rule-in	0.81
	Neutral	5—10	19	28	1.39	Indeterminate	
	Negative	11	1	15	0.14	Moderate	
15 months (n = 64)	Positive	<8	10	0	∞	SpPin, rule-in	0.96
	Neutral	8—11	5	15	1.09	Indeterminate	
	Negative	12	0	34	0	SnNout, Rule-out	
18 months (n = 98)	Positive	<10	20	0	∞	SpPin, rule-in	0.84
	Neutral	10—12	21	18	1.22	Indeterminate	
	Negative	13	7	32	0.23	Small	
24 months (n = 73)	Positive	<8	14	0	∞	SpPin, rule-in	0.90
	Neutral	8—10	13	22	0.85	Indeterminate	
	Negative	11	3	21	0.20	Moderate	
30 months (n = 82)	Positive	<9	20	0	∞	SpPin, rule-in	0.86
	Neutral	9—11	16	12	1.21	Indeterminate	
	Negative	12	7	27	0.24	Small	

DD=developmental delays; DT=developing typically; AUC = area under the receiver operating characteristics curve.

procedure similar to that used in a medical setting and proposed multiple cutoff points for clinical application for different age groups from the empirical data.

A comparison of the LR+s in Tables 3 and 4 shows that those with positive test results in multiple cutoff strategies were always much higher than those in a single cutoff strategy. Similarly, LR-s of those with negative test results in multiple cutoff points usually were equal to or lower than those in single cutoff strategy.

The DOR of the 15-month checklist showed the best results of differentiating positive and negative results. Therefore, using multiple cutoff strategies should be able to decrease false positive or false negative results. However, there were some checklists with an LR that had a small diagnostic impact, and a large proportion of children held a score in the indeterminate range of diagnostic impact. A high false negative rate could be expected if no second screening tests are provided. Therefore, the periodical developmental surveillance procedure as suggested by the National Health Bureau of the Department of Health and multisource information collection in healthy baby clinics are very important.

From June 2010, the developmental surveillance of each child has been conducted by using the developmental surveillance items in the Child Health Pamphlet (DICHP)[12] at ages 1, 2—4, 4—10, 10—18, 18—24, 24—36, and 36—84 months. The results of this study provided psychometric information for the Taipei II for children age 5 months to 36 months, and it supports the screening decisions in Taiwan.

However, if the policy used the DICHP as the first screening measure, then the psychometric properties of the Taipei II as the second screening measure might need further exploration. Besides, among the suggested seven age stages proposed by the Department of Health, two ages, 24—36 and 36—84 months, are recommended as the most important stages. The psychometric properties of Taipei II checklists for older age groups need further study as well.

Although the test—retest reliability coefficient of the 30-month checklist was not significant, the test—retest agreement of each individual item was above 90%. Further analysis with a contingency table found that the agreement of the total score was 79%. One child who was rated 10 on the total score at the first test, got a 12 on the second test. The two items that changed from a rating of failure to a rating of success in the two repeated tests were "articulation of speech not clear enough to be understood by the closest adults" and "very uncooperative during screening process, exhibits one of the following behaviors: (1) not listening to the instruction, not looking at the demonstration procedure; (2) not looking at and following the caregiver's pointing direction; (3) not willing to point and show; (4) grabbing things from the adults and playing by himself; (5) seemingly unable to understand instruction." Those behaviors might be context-dependent and more subjective, and at times, they may be easily rated differently. Besides, the score range of this 30-month group was narrow, from 10 to 12 at the first test, and from 11 to 12 at the second test. The small variability might also cause low or nonsignificant correlation coefficients.[4]

The previous unpublished study showed that for children age 2 years to 3.5 years, the test—retest reliability of the total scores of the Taipei II by a well-trained tester were 0.71 ($p < 0.001$) (Cheng, personal communication). Therefore, the test—retest reliability of the 30-month checklist needs further study.

All of the three poor agreement items were negative statement items. The third item of the 6-month checklist is "using forearm support to raise the upper body and turn the head freely in the tummy position (not move head abruptly and not always hyperextend the neck)." Five children who were rated as passing this item at the first test, were rated as failing at the retest, and four children were rated reversely. The "move head abruptly" and "not always hyperextend the neck" might be not easily understood by parents. The third item of the 12-month checklist is a marked item: "Can only play things by mouthing or throwing, no other ways (such as shaking, squeezing, knocking, pulling, etc)." One child who was rated as passing at the first test was rated as failing at the retest, and four children were rated reversely. These parents might not fully understand the whole statement, and perhaps focused on the first part or the last part of the sentences. The ninth item of the 24-month checklist is "unable to imitate a single phrase due to: (1) no motivation to imitate sound; or (2) articulation too poor to be understood." This statement contains one unachieved skill statement and two why statements. This type of presentation also might mislead the parents' rating behavior. Therefore, we suggest revising these three items in the future.

The original Taipei II has items marked to be weighted more when the tester interprets the test results. Before calculating the total score in order to analyze multilevel LRs and ROCs, the authors calculated the sensitivity, specificity, and YI for each item at the beginning of this study. We found that most marked items had high specificities, $> 90\%$, but they were not necessarily higher than other items in the same checklist. And the YIs of the marked items were not all higher than nonmarked items. Therefore, we decided to weigh each item equally and added the passing item number to the total score. Further study using the Rasch model to analyze the difficulty level and discrimination ability of each item is recommended.

The mean ages, proportion of sexes and prematurity of the children, and the proportion of career mothers were significantly different between the DD and DT groups in this study. The convenient sample of this study was the main reason. However, we examined psychometric properties in individual age-appropriate checklists, and the higher mean age in the whole DD group did not cause bias. Premature infants usually have an increased risk of poor motor, mental, and behavior-related developmental outcomes.[22,23] The higher proportion of prematurity in the DD group was expected and did not influence the results of this study. The Taiwanese norm of a developmental test shows that there are no significant differences in developmental scores between boys and girls at the same age level.[24] The different proportion of males between the two groups did not influence the psychometric properties of the Taipei II of this study. Because the psychometric properties of the Taipei II were examined on individual age-appropriate checklists, we further analyzed the differences of the career mother

proportions at seven age levels between two groups. The results showed that only career mother proportions at 18-month and 30-month age levels were significantly different. There were 27 career mothers (54%) in the DT group and only 11 career mothers (23%) in the DD group at the 18-month level. We further analyzed the AUC and multiple likelihood ratio for toddlers with career mothers and those whose mothers were not career mothers, and found that the values of AUC were all above 0.76 ($p < 0.05$). When the value of the total score 10 was chosen as the cutoff point of the positive test result, the diagnostic impact still met the SpPin criteria in the separate groups. A similar procedure was done at the 30-month age level, and the results were similar. In summary, the differences of the demographic data in the two groups did not have a large impact on the results of this study.

The limitations of the study were as follows. First, due to the low prevalence of children with developmental delays, collecting enough children with developmental delays from a consecutive sample was difficult. We then used the case control study to examine the psychometrics of the Taipei II. The diagnostic accuracy obtained from a case control study is usually higher than that from a consecutive sample study because the cases in the case control study are usually severer than average cases.[25] However, more than 1% of children in the DD group passed all the items in each checklist in this study. The cases of this study might not all be severe ones. Second, the sample of this study mostly came from the northern area and the mothers' educational level was higher than high school graduation. The psychometric properties of other populations in different areas and different cultural backgrounds need further study. Third, the Taipei II is currently encouraged to be used for children up to 72 months of age. The children in this study were age 5.5 months to 35.5 months; hence, the result is not applicable to preschoolers older than 35.5 months. The multiple likelihood ratios of Taipei II for children older than 3 years need further study. Fourth, the children developing typically were judged by the pediatricians without diagnostic developmental tests in this study. Children with mild or borderline emotional/behavioral developmental problems placed into the normal group might be misclassified.[26] However, the developmental surveillance items of the Child Health Pamphlet, a developmental test with acceptable reliability and validity, [12] has been used to increase the diagnostic accuracy of children developing typically by the pediatricians in this study.

Acknowledgments

We sincerely appreciate the participating dyads and the physical therapist, Yu-Ling Kuo, for her help with data collection. A special thanks goes to the National Science Council, R.O.C. (Taiwan) and Department of Health, for this study was supported by them (Grant NSC 96-2314-B-002-074-MY3 and DOH95-HP-1205). We would also like to thank the following departments that helped us to collect data: Pediatric Department and Rehabilitation Department of the National Taiwan University Hospital; Department of Rehabilitation Medicine, Taipei Veterans General Hospital; Department of Rehabilitation Medicine, Shin-Kong Wu Ho-Su Memorial Hospital; Department of Child Psychology, Taipei Municipal Women's and Child's Hospital; Social Welfare Department of Taipei City; Chuang- Kai-Chuan Pediatric Clinic in Chiayi City.

References

1. Spittle AJ, Orton J, Doyle LW, Boyd R. Early developmental intervention programs post hospital discharge to prevent motor and cognitive impairments in preterm infants. *Cochrane Database Syst Rev* 2007. CD005495.
2. Washington K, Scott DT, Johnson KA, Wendel S, Hay AE. The Bayley Scales of Infant Development-II and children with developmental delays: a clinical perspective. *J Dev Behav Pediatr* 1998;**19**:346—9.
3. Sices L, Feudtner C, McLaughlin J, Drotar D, Williams M. How do primary care physicians manage children with possible developmental delays? A national survey with an experimental design. *Pediatrics* 2004;**113**:274—82.
4. Anatasi A, Urbina S. *Psychological testing*. 7th ed. Upper Saddle River, NJ: Prentice Hall; 1997.
5. Brenneman SK. Assessment and testing of infant and child development. In: Tecklin JS, editor. *Pediatric physical therapy*. 3rd ed. Philadelphia: JB Lippincott; 1999. p. 28—70.
6. Liao HF, Yang MC, Cheng LY, Hsieh WS. The cost of the two cutoff strategies of the Taipei City Developmental Screening Checklist for Preschoolers 2nd version. *Formosan J Med* 2009; **13**:9—22. in Chinese.
7. Liao HF, Cheng LY, Hsieh WS, Yang MC. Selecting a better cut-off strategy of a developmental screening test based on overall diagnostic indexes and the total expected utilities of professional preferences. *J Formos Med Assoc* 2010;**109**: 209—18.
8. Skellern CY, Rogers Y, O'Callaghan MJ. A parent-completed developmental questionnaire: follow up of ex-premature infants. *J Paediatr Child Health* 2001;**37**:125—9.
9. Elbers J, Macnab A, McLeod E, Gagnon F. The ages and stages questionnaires: feasibility of use as a screening tool for children in Canada. *Can J Rural Med* 2008;**13**:9—14.
10. Barry HC, Ebell MH. Test characteristics and decision rules. *Endocrinol Metab Clin North Am* 1997;**26**:45—65.
11. Straus SE, Richardson WS, Glaszion P, Haynes RB. *Evidence-based medicine: how to practice and teach EBM*. 3rd ed. London, England: Elsevier Churchill Livingstone; 2005.
12. Liao HF, Cheng LY, Hsieh WS, Yang MC, Tsou KS, Tsai KY. The reliability and validity of the developmental surveillance items of Child Health Pamphlet. *Formosan J Med* 2008;**12**:502—12. in Chinese.
13. Glas AS, Lijmer JF, Prins MH, Bonsel GJ, Bossuyt PMM. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol* 2003;**56**:1129—35.
14. Glascoe FP. Parents' concerns about children's development: prescreening technique or screening test? *Pediatrics* 1997;**99**: 522—8.
15. Aylward G. Conceptual issues in developmental screening and assessment. *J Dev Behav Pediatr* 1997;**18**:340—9.
16. Pewsner D, Battaglia M, Minder C, Marx A, Bucher HC, Egger M. Ruling a diagnosis in or out with "SpPIn" and "SnNOut": a note of caution. *BMJ* 2004;**329**:209—13.
17. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature, III: how to use an article about a diagnostic test, B: what are the results and will they help me in caring for my patients? *JAMA* 1994;**271**:703—7.
18. Fischer JE, Bachmann LM, Jaeschke R. A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. *Intensive Care Med* 2003;**29**:1043—51.

19. Park SH, Goo JM, Jo CH. Receiver operating characteristic (ROC) curve: practical review for radiologists. *Korean J Radio* 2004;**5**:11—8.
20. Hunink M, Glasziou P, Siegel J, Weeks J, Pliskin J, Elstein A, et al. *Decision making in health and medicine: integrating evidence and values*. UK, Cambridge: University Press; 2001.
21. Baxter P. Normality and abnormality. *Dev Med Child Neurol* 2006;**48**:867.
22. Chen PS, Jeng SF, Tsou KI. Developmental function of very-low-birth-weight infants and full-term infants in early childhood. *J Formos Med Assoc* 2004;**103**:23—31.
23. Liao HF, Wang TM, Yao G, Lee WC. Concurrent validity of Comprehensive Developmental Inventory for Infants and Toddlers with Bayley Scales of Infant Development-II in preterm infants. *J Formos Med Assoc* 2005;**104**: 731—7.
24. Wang TM, Su CW, Liao HF, Lin LY, Chou KS, Lin AH. The standardization of the Comprehensive Developmental Inventory for Infants and Toddlers. *Psychol Testing* 1998;**45**:19—46. in Chinese.
25. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;**282**:1061—6.
26. Lavigne JV, Binns HJ, Christoffel KK, Rosenbaum D, Arend R, Smith K, et al. Behavioral and emotional problems among preschool children in pediatric primary care: prevalence and pediatricians' recognition. *Pediatrics* 1993; **91**:649—55.